Taylor, A. E., Ware, J. J., Gage, S. H., Smith, G. D., & Munafò, M. R. (2015). Using molecular genetic information to infer causality in observational data: Mendelian randomisation. *Current Opinion in Behavioral Sciences*, *2*, 39-45. https://doi.org/10.1016/j.cobeha.2014.08.002

Using Genetic Information to Infer Causality in Observational Data: Mendelian Randomization

Amy E. Taylor [1,2], Jennifer J. Ware [1,2], Suzanne H. Gage [1,2], George Davey Smith [1], Marcus R. Munafò [1,2]

1. MRC Integrative Epidemiology Unit (IEU) at the University of Bristol, United Kingdom.

2. UK Centre for Tobacco and Alcohol Studies and School of Experimental Psychology, University of Bristol, United Kingdom

Corresponding author: Marcus R. Munafò, School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, United Kingdom. T: +44.117.9546841; F: +44.117.9288588; E: marcus.munafo@bristol.ac.uk

Abstract

Determining whether associations between lifestyle behaviours and health outcomes are causal is difficult in observational data. However, as genetic variants associated with these behaviours are discovered, this will provide opportunities for testing causality using Mendelian randomization methods. These use genetic variants as proxies for exposures to minimise biases associated with observational data, enabling stronger causal inference. Here we review the principles and main approaches for conducting Mendelian randomization studies, and discuss recent methodological developments for investigating more complex causal pathways. Mendelian randomization offers considerable promise for improving our understanding of the causal relationships between lifestyle behaviours and health outcomes, and its application will increase as more genetic variants robustly associated with behavioural phenotypes are identified.

Keywords: Mendelian randomization, causality, health behaviours

Using Genetic Information to Infer Causality in Observational Data:

Mendelian Randomization

Genome-wide association studies (GWAS) are revealing genetic variants associated with phenotypes such as tobacco use [1-3], obesity [4] and educational attainment [5]. These findings have advanced our understanding of the neurobiological basis of these phenotypes [6], but also offer the opportunity to use this information to make causal inferences regarding their effects on a range of outcomes. Mendelian randomization (MR) is based on instrumental variable (IV) methods developed in the economics literature, and aims to minimise problems of measurement bias, confounding and reverse causality intrinsic to observational studies. IV analysis requires a variable that is a proxy or *instrument* for the exposure of interest, which must meet a number of criteria: 1) association with the exposure of interest; 2) no association with the outcome of interest, apart from via the exposure; 3) no association with confounders affecting the relationship under investigation; and, 4) unable to introduce potential confounding in to the relationship [7]. Given an appropriate instrument, confounders will be randomly distributed across the conditions of interest in the same way as a randomized trial – (see Figure 1). This is particularly important in observational studies; confounders may be difficult to adequately adjust for, and some may be impossible to measure or unknown [8]. An ideal instrument would be unrelated to measured or unmeasured confounders, known or unknown.

Insert Figure 1 about here.

Mendelian Randomization

Mendelian randomization uses genetic variants as instruments for environmental exposures [9,10]. These can take the form of individual single nucleotide polymorphisms (SNPs), or polygenic risk scores, which must be robustly associated with the exposure of interest (e.g., smoking heaviness or alcohol use) (see Figure 2). The principle of MR relies

on the basic (but approximate) laws of Mendelian genetics (segregation and independent assortment). If these hold then, at a population level, genetic variants will not be associated with potential confounders [11,12]. The SNP or risk score must also not directly affect the outcome being investigated. Certain exposures, such as number of cigarettes or amount of alcohol consumed, allow for this assumption to be tested, as the effect of gene on the outcome can be assessed in those unexposed to the putative causal risk factor. For example, if a gene meant to be a proxy for number of cigarettes smoked has a relationship with an outcome in those who have never smoked, this suggests a direct effect of the gene.

Insert Figure 2 about here.

SNPs or risk scores have other potential benefits over observational studies. For example, genes act on exposures over a long period, and therefore better index long-term environmental exposure than self-report measures taken at a specific time point. Also, MR effectively rules out reverse causation: the outcome cannot affect genotype. Therefore, if specific genetic variants associated with environmental exposures are identified, it may be possible to use MR to explore the causal effects of those exposures. Where variants have been identified, MR studies have already been undertaken, for example looking at the effects of alcohol use [13,14] and tobacco use [15-18]. These have provided evidence that maternal alcohol drinking in pregnancy adversely impacts offspring educational outcomes [13], that alcohol consumption increases blood pressure and body mass index (BMI) [14], that smoking lowers BMI [15], and that maternal smoking in pregnancy reduces offspring birth weight [18].

MR can enable causal inference in two broad ways (See Figure 3). First, a direct association between a genetic instrument and the outcome of interest can provide evidence for the *existence* of a causal relationship between exposure and outcome. Second, the magnitude of the association between a genetic variant and the exposure, and between the genetic variant and outcome, can be used to estimate the *magnitude* of the causal effect of

the exposure on the outcome (using methods such as two stage least squares regression). As genotype will affect exposure over a lifetime, MR can in principle allow for more accurate estimation of the magnitude of a causal effect than a direct assessment taken at a single time point [19] although for the same reason it may over-estimate the likely magnitude of an intervention effect. For example, an intervention delivered in middle age will only partially reduce the lifetime exposure to a risk factor that is estimated from MR analyses.

Insert Figure 3 about here.

Two Sample Mendelian Randomization

Commonly, the association between a genetic variant and the exposure, and between the genetic variant and the outcome, are estimated in the same sample. However, this may not always be possible if exposure and outcomes are not measured in the same samples, or if the exposure has only been measured in a subset of the total sample [20]. In two sample MR, the genotype-exposure and genotype-outcome associations are estimated in different samples and these estimates then combined to provide an estimate of the causal exposure-outcome association [21]. As both of these parameters are estimates, the standard error of the exposure-outcome association needs to be adjusted using appropriate methods [20]. Two sample MR does not usually lead to a substantial loss of statistical power [21], so this type of design may be a more cost effective approach [20].

Two-Step Mendelian Randomization

Establishing that an association is causal is valuable in itself, but of potentially greater interest is establishing the mechanism through which this causal association operates. It may be possible to investigate causal mechanisms between an exposure and an outcome using a two-step MR approach [22]. This type of analysis requires a genetic variant which associates with the exposure of interest and a separate genetic variant which associates with the mediating factor of interest. For example, there is growing interest in the

role of epigenetic mediators of environmental exposures, but epigenetic markers (as with any other biomarker) are vulnerable to confounding and reverse causality. Here, a genetic proxy for the exposure of interest is used to assess the causal relationship between the environmental exposure and a potential mediator such as methylation (step 1, see Figure 4A). Next, a genetic proxy for the mediator (here, DNA methylation) is used to interrogate the causal relationship between the mediator and the outcome of interest (step 2, see Figure 4B). This approach enables a triangulation of evidence to infer a mediating role for, in this case, methylation in the causal pathway between the environmental exposure and the outcome of interest. It can in principle be applied to other potential mediators (e.g., metabolite levels).

Insert Figure 4 about here.

Bidirectional and Network Mendelian Randomization

Early MR studies focused on a single direction of causality, such as the effects of alcohol consumption on cardiovascular risk [14], but in many cases the relationship may be bidirectional. For example, tobacco use has been shown to lower BMI [15], but BMI may also affect smoking behaviour if individuals smoke in order to control their weight. In cases such as this, where genetic instruments for both the exposure and the outcome are available, MR analysis may be performed in both directions. Bidirectional MR has been used previously to investigate the direction of causality between BMI and a number of other factors, including vitamin D and C-reactive protein levels [23,24]. A more complex problem arises when multiple phenotypes that may influence each other in a causal network are considered. Methods are currently being developed, using multiple genetic variants, which allow assessment of causal directions in pathways with correlated phenotypes [20, 25, 26].

Limitations to Mendelian Randomization

MR studies require much larger sample sizes than conventional exposure-outcome analyses. As a general rule, sample sizes for MR studies can be calculated by multiplying the required observational sample size by the inverse of the variance (R^2 or square of the correlation coefficient) in the exposure of interest explained by the genetic instrument. For example, for a genetic variant explaining 1% of the variance in an exposure, the sample size would need to be 100 times greater than the sample size required to detect the true causal effect between the directly measured exposure and the outcome. Statistical code and online calculators are now available for determination of sample sizes required for MR studies for both continuous and categorical outcomes [27-29]. Whilst collaborative consortia (see Text Box 1) offer a potential solution to the issue of power in MR studies, combining phenotypic outcomes across many different studies can be challenging, particularly for behavioural exposures and outcomes.

Insert Text Box 1 about here.

It is also only possible to use MR to study the effects of exposures for which genetic variants have been identified. Whilst GWAS have been successful in identifying variants that influence a number of traits, there are still many exposures for which we do not yet have suitable instruments. In addition, genetic variants may be population-specific and not suitable for use in all ancestral groups. For example, a variant in the *ALDH2* gene, which strongly influences alcohol consumption, is used in MR studies in East Asian populations, but occurs at too low a frequency for use in MR studies in European populations [30]. Critically, genetic variants in MR studies must be associated with the exposure of interest within the analysis sample *and* must show robust evidence for association with the same exposure in independent samples. Performing MR analyses using genetic instruments that have been discovered within the analysis sample but have not been independently replicated can lead to causal inference in the absence of true causal effects, because associations between genetic variants and exposures may just be chance findings. In

addition, as effect sizes between genetic variants and phenotypes are often inflated in discovery samples (also known as the Beavis effect or Winner's Curse), performing MR analyses within discovery samples can result in biased causal effect sizes [31].

Biased estimates of effect sizes may also be obtained if the measured exposure does not fully capture the causal exposure through which the genetic variant operates [31]. For example, a variant in the nicotinic receptor alpha-5 subunit protein, rs16969968, influences lifetime tobacco exposure, but this is not well captured by self-report measures of smoking (e.g., cigarettes per day). MR of lung cancer data using cigarettes per day as the intermediate variable indicates a causal odds ratio for lung cancer of 2,180 per pack of cigarettes smoked per day, compared to only 2.6 from observational analysis [32]. In contrast, using cotinine, a metabolite of nicotine and a more precise objective measure of tobacco exposure, produce effect sizes which are more consistent with observational findings [33]. In the absence of appropriate intermediate exposure measures, MR can still be used to infer causality, but it may not be possible to accurately estimate causal magnitudes of effect. Furthermore, MR studies can be informative about the effects of lifelong exposure to a risk factor, but are usually not appropriate for investigating the impact of short-term changes in risk factors on health outcomes. MR studies will also rarely provide information about the mechanisms underlying a causal relationship (although two-step MR can provide this).

Whilst MR can minimise many of the biases associated with conventional epidemiological studies, there are ways in which MR can still be confounded. Spurious associations between genes and outcomes may arise through population stratification if samples are made up of populations of more than one ancestry, which have different allele frequencies and different levels of disease outcomes [19]. Therefore, care should be taken to identify and appropriately control for genetic ancestry. Confounding may also arise if the variant has pleiotropic effects which influence the outcome other than through the exposure of interest, or if the variant is in linkage disequilibrium with another genetic variant which also influences the outcome [20]. In such cases, one cannot be confident that any "causal" effect

observed operates through the exposure of interest. In some MR studies of lifestyle behaviours, it may be possible to perform a test of pleiotropy by investigating associations of the genetic variant with the outcome in individuals not exposed to the behaviour. This has been demonstrated in MR studies of alcohol use in East Asians, which have stratified analyses by sex. The alcohol-related variant influences blood pressure in males (who consume alcohol) but not in females (who tend not to consume alcohol in many East Asian cultures for social and historical reasons), indicating that the likely mechanism of the genetic effect on blood pressure is through alcohol consumption [34]. However, whilst stratifying on an exogenous variable such as sex, as described above, can be a useful tool in some MR studies, care must be taken not to reintroduce confounding through collider bias [35,36]. This can occur when MR analyses are stratified on the measured exposure of interest and can amplify or mask associations between the genetic variant and outcome within the exposure strata [37].

A further potential concern is the possibility of canalization, which is the process of developmental compensation to buffer against the effects of disruptive genetic or environmental influences during development [9]. If exposure to elevated levels of a risk factor during foetal development or post-natal growth results in tissue changes which compensate for this, the genetic variant will still associate with the risk factor of interest, but any potential effects on a disease outcome may be reduced. However, canalization is less problematic for exposures which tend to occur later in development, such as smoking and alcohol consumption [7].

There are a number of other statistical issues in relation to MR, particularly surrounding the use of two-stage instrumental variable analysis (e.g., weak instrument bias). These are beyond the scope of this review, but are discussed in detail elsewhere [38-40].

Conclusions

Inferring causation from observational data is notoriously problematic. Although MR relies on certain assumptions that may not always apply, it nevertheless has the potential to

dramatically advance our understanding of the causal role of modifiable environmental exposures on a variety of outcomes. As genome-wide association studies continue to reveal variants associated with a range of behavioural phenotypes, the applications of MR will grow. In particular, risk scores that capture a substantial proportion of the phenotypic variation in behavioural outcomes will enable us to apply MR more extensively, by providing stronger instruments. Genome-wide association studies have enjoyed substantial success in many areas, and are beginning to realise similar success for other phenotypes (e.g., psychiatric outcomes such as schizophrenia) where understanding the causal role of these phenotypes will be of considerable scientific and societal importance.

Acknowledgements

The authors are members of the UK Centre for Tobacco and Alcohol Studies, a UKCRC Public Health Research: Centre of Excellence. Funding from British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. This work was supported by the Medical Research Council (grant number MC_UU_12013/1-9).

References

1. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, Berrettini W, Knouff CW, Yuan X, Waeber G, et al.: **Meta-analysis and imputation refines the association of 15q25 with smoking quantity**. *Nat Genet* 2010, **42**:436-U475.

2. Furberg H, Kim Y, Dackor J, Boerwinkle E, Franceschini N, Ardissino D, Bernardinelli L, Mannucci P, Mauri F: **Genome-wide meta-analyses identify multiple loci associated with smoking behavior**. *Nat Genet* 2010, **42**:441-447.

3. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, Sulem P, Rafnar T, Esko T, Walter S, et al.: **Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior**. *Nat Genet* 2010, **42**:448-453.

4. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Lango Allen H, Lindgren CM, Luan J, Magi R, et al.: **Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index**. *Nat Genet* 2010, **42**:937-948.

5. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, Westra HJ, Shakhbazov K, Abdellaoui A, Agrawal A, et al.: **GWAS of 126,559 individuals identifies genetic variants associated with educational attainment**. *Science* 2013, **340**:1467-1471.

6. Fowler CD, Lu Q, Johnson PM, Marks MJ, Kenny PJ: **Habenular alpha5 nicotinic receptor subunit signalling controls nicotine intake**. *Nature* 2011, **471**:597-601.

7. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G: **Mendelian randomization: using genes as instruments for making causal inferences in epidemiology**. *Stat Med* 2008, **27**:1133-1163.

8. Tatsioni A, Bonitsis NG, Ioannidis JP: **Persistence of contradicted claims in the literature**. *JAMA* 2007, **298**:2517-2526.

** 9. Davey Smith G, Ebrahim S: **'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?** *Int J Epidemiol* 2003, **32**:1-22.

Seminal paper describing the basic principles of Mendelian randomization, and how they can be applied to improve our understanding of the environmental causes of disease.

10. Davey Smith G, Ebrahim S: **What can mendelian randomisation tell us about modifiable behavioural and environmental exposures?** *BMJ* 2005, **330**:1076-1079.

11. Davey Smith G: **Mendelian randomization for strengthening causal inference in observational studies: Application to gene x environment interactions**. *Perspect Psychol Sci* 2010, **5**:527-545.

12. Davey Smith G, Lawlor DA, Harbord R, Timpson N, Day I, Ebrahim S: **Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology**. *PLoS Med* 2007, **4**:e352.

13. Zuccolo L, Lewis SJ, Davey Smith G, Sayal K, Draper ES, Fraser R, Barrow M, Alati R, Ring S, Macleod J, et al.: **Prenatal alcohol exposure and offspring cognition and school performance. A 'Mendelian randomization' natural experiment**. *Int J Epidemiol* 2013, **42**:1358-1370.

14. Lawlor DA, Nordestgaard BG, Benn M, Zuccolo L, Tybjaerg-Hansen A, Davey Smith G: **Exploring causal associations between alcohol and coronary heart disease risk factors: findings from a Mendelian randomization study in the Copenhagen General Population Study**. *Eur Heart J* 2013, **34**:2519-2528.

15. Freathy RM, Kazeem GR, Morris RW, Johnson PC, Paternoster L, Ebrahim S, Hattersley AT, Hill A, Hingorani AD, Holst C, et al.: **Genetic variation at CHRNA5-CHRNA3-CHRNB4 interacts with smoking status to influence body mass index**. *Int J Epidemiol* 2011, **40**:1617-1628.

16. Lewis SJ, Araya R, Davey Smith G, Freathy R, Gunnell D, Palmer T, Munafo M: **Smoking is associated with, but does not cause, depressed mood in pregnancy--a mendelian randomization study**. *PLoS One* 2011, **6**:e21689.

17. Bjorngaard JH, Gunnell D, Elvestad MB, Davey Smith G, Skorpen F, Krokan H, Vatten L, Romundstad P: **The causal role of smoking in anxiety and depression: a Mendelian randomization analysis of the HUNT study**. *Psychol Med* 2013, **43**:711-719.

18. Tyrrell J, Huikari V, Christie JT, Cavadino A, Bakker R, Brion MJ, Geller F, Paternoster L, Myhre R, Potter C, et al.: **Genetic variation in the 15q25 nicotinic acetylcholine receptor gene cluster (CHRNA5-CHRNA3-CHRNB4) interacts with maternal self-reported smoking status during pregnancy to influence birth weight**. *Hum Mol Genet* 2012, **21**:5344-5358.

19. Davey Smith G, Ebrahim S: **Mendelian randomization: prospects, potentials, and limitations**. *Int J Epidemiol* 2004, **33**:30-42.

** 20. Davey Smith G, Hemani G: **Mendelian randomization: Genetic anchors for causal inference in epidemiological studies.** *Hum Mol Genet* 2014.

Updated review of the basic principles of Mendelian randomization, incorporating recent methodological developments such as bidirectional and two-step Mendelian randomization.

* 21. Pierce BL, Burgess S: **Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators**. *Am J Epidemiol* 2013, **178**:1177-1184.

Uses simulated data to illustrate the potential efficiency of two sample instrumental variable designs.

22. Relton CL, Davey Smith G: **Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease**. *Int J Epidemiol* 2012, **41**:161-176.

23. Vimaleswaran KS, Berry DJ, Lu C, Tikkanen E, Pilz S, Hiraki LT, Cooper JD, Dastani Z, Li R, Houston DK, et al.: **Causal relationship between obesity and vitamin D status: bi-directional Mendelian randomization analysis of multiple cohorts**. *PLoS Med* 2013, **10**:e1001383.

24. Timpson NJ, Nordestgaard BG, Harbord RM, Zacho J, Frayling TM, Tybjaerg-Hansen A, Davey Smith G: **C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization**. *Int J Obes (Lond)* 2011, **35**:300-308.

25. Peng CH, Jiang YZ, Tai AS, Liu CB, Peng SC, Liao CT, Yen TC, Hsieh WP: **Causal inference of gene regulation with subnetwork assembly from genetical genomics data**. *Nucleic Acids Res* 2014, **42**:2803-2819.

26. Neto EC, Keller MP, Attie AD, Yandell BS: **Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes**. *Ann Appl Stat* 2010, **4**:320-339.

* 27. Freeman G, Cowling BJ, Schooling CM: **Power and sample size calculations for Mendelian randomization studies using one genetic instrument**. *Int J Epidemiol* 2013, **42**:1157-1163.

Useful guide to sample size and power calculations in Mendelian randomisation designs, in cases where a single genetic instrument is used. Provides formulas for sample size and power calculations.

28. Brion MJ, Shakhbazov K, Visscher PM: **Calculating statistical power in Mendelian randomization studies**. *Int J Epidemiol* 2013, **42**:1497-1501.

29. Burgess S: **Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome**. *Int J Epidemiol* 2014.

30. Lewis SJ: **Mendelian randomization as applied to coronary heart disease, including recent advances incorporating new technology**. *Circ Cardiovasc Genet* 2010, **3**:109-117.

31. Taylor AE, Davies NM, Ware JJ, Vanderweele T, Davey Smith G, Munafo MR: **Mendelian randomization in health research: Using appropriate genetic variants and avoiding biased estimates**. *Econ Hum Biol* 2014, **13**:99-106.

32. VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P: **Methodological challenges in mendelian randomization**. *Epidemiology* 2014, **25**:427-435.

33. Munafo MR, Timofeeva MN, Morris RW, Prieto-Merino D, Sattar N, Brennan P, Johnstone EC, Relton C, Johnson PC, Walther D, et al.: **Association between genetic variants on chromosome 15q25 locus and objective measures of tobacco exposure**. *J Natl Cancer Inst* 2012, **104**:740-748.

* 34. Chen L, Davey Smith G, Harbord RM, Lewis SJ: **Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach**. *PLoS Med* 2008, **5**:e52.

Elegant example of the application of Mendelian randomization techniques to the question of whether alcohol intake is causally associated with higher blood pressure.

* 35. Glymour MM, Tchetgen EJ, Robins JM: **Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions**. *Am J Epidemiol* 2012, **175**:332-339.

Clear discussion of the assumptions of the instrumental variable approach and methods to assess the validity of these assumptions in Mendelian randomization studies. Includes a description of collider bias.

36. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, Poole C: **Illustrating bias due to conditioning on a collider**. *Int J Epidemiol* 2010, **39**:417-420.

37. Gage SH, Davey Smith G, Zammit S, Hickman M, Munafo MR: **Using mendelian randomisation to infer causality in depression and anxiety research**. *Depress Anxiety* 2013, **30**:1185-1193.

38. Burgess S, Thompson SG: **Bias in causal estimates from Mendelian randomization studies with weak instruments**. *Stat Med* 2011, **30**:1312-1323.

39. Burgess S, Thompson SG, CRP-CHD-Genetics-Collaboration, Burgess S, Thompson SG, Andrews G, Samani NJ, Hall A, Whincup P, Morris R, et al.: **Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables**. *Stat Med* 2010, **29**:1298-1311.

40. Harbord RM, Didelez V, Palmer TM, Meng S, Sterne JA, Sheehan NA: **Severity of bias of a simple estimator of the causal odds ratio in Mendelian randomization studies**. *Stat Med* 2013, **32**:1246-1258.

41. Thorgeirsson TE, Geller F, Sulem P, Rafnar T, Wiste A, Magnusson KP, Manolescu A, Thorleifsson G, Stefansson H, Ingason A, et al.: **A variant associated with nicotine dependence, lung cancer and peripheral arterial disease**. *Nature* 2008, **452**:638-642.

42. Ware JJ, van den Bree MB, Munafo MR: **Association of the CHRNA5-A3-B4 gene cluster with heaviness of smoking: a meta-analysis**. *Nicotine Tob Res* 2011, **13**:1167-1175.

Figure 1. Randomization by Intervention and Genetics.

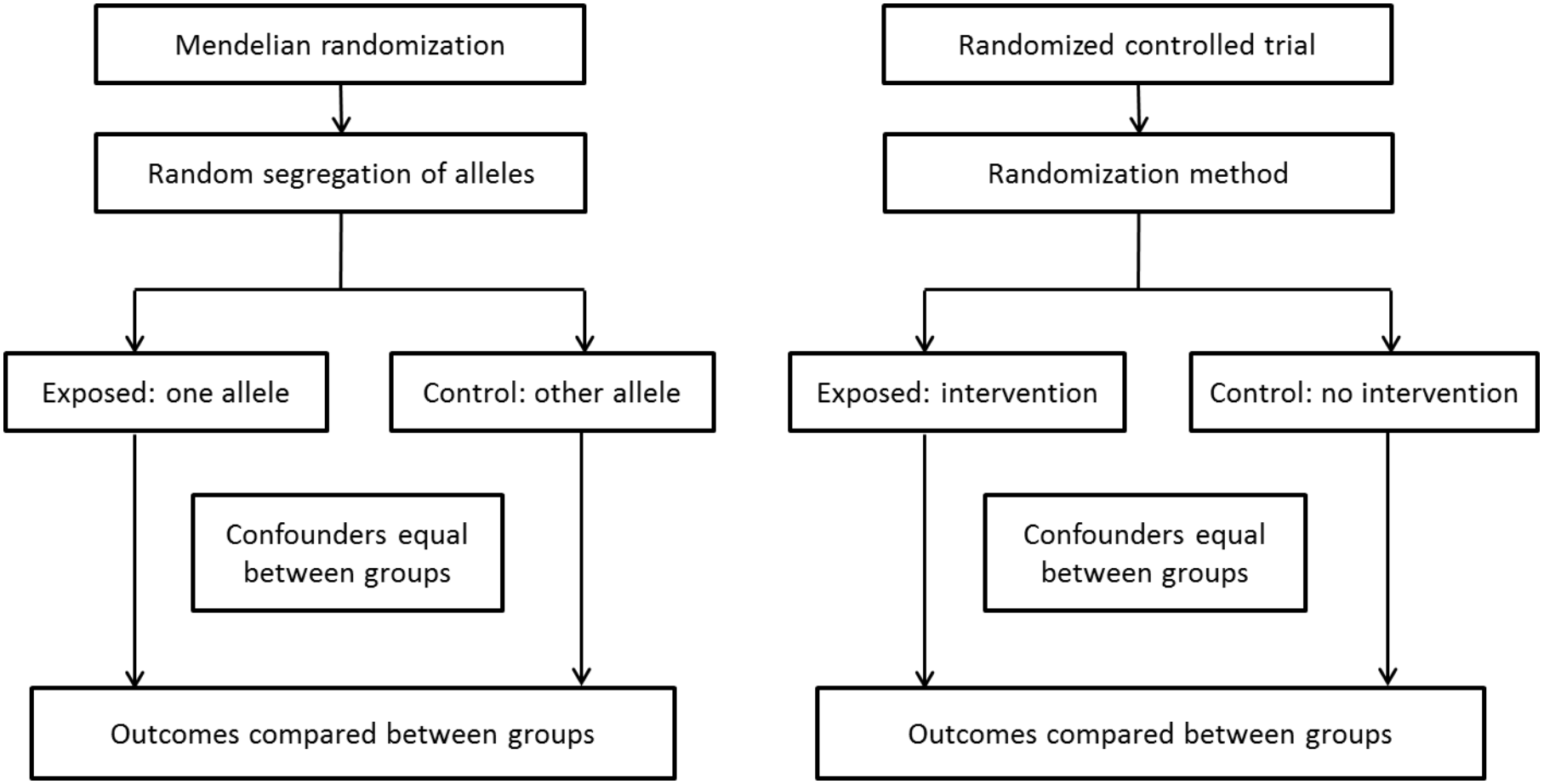
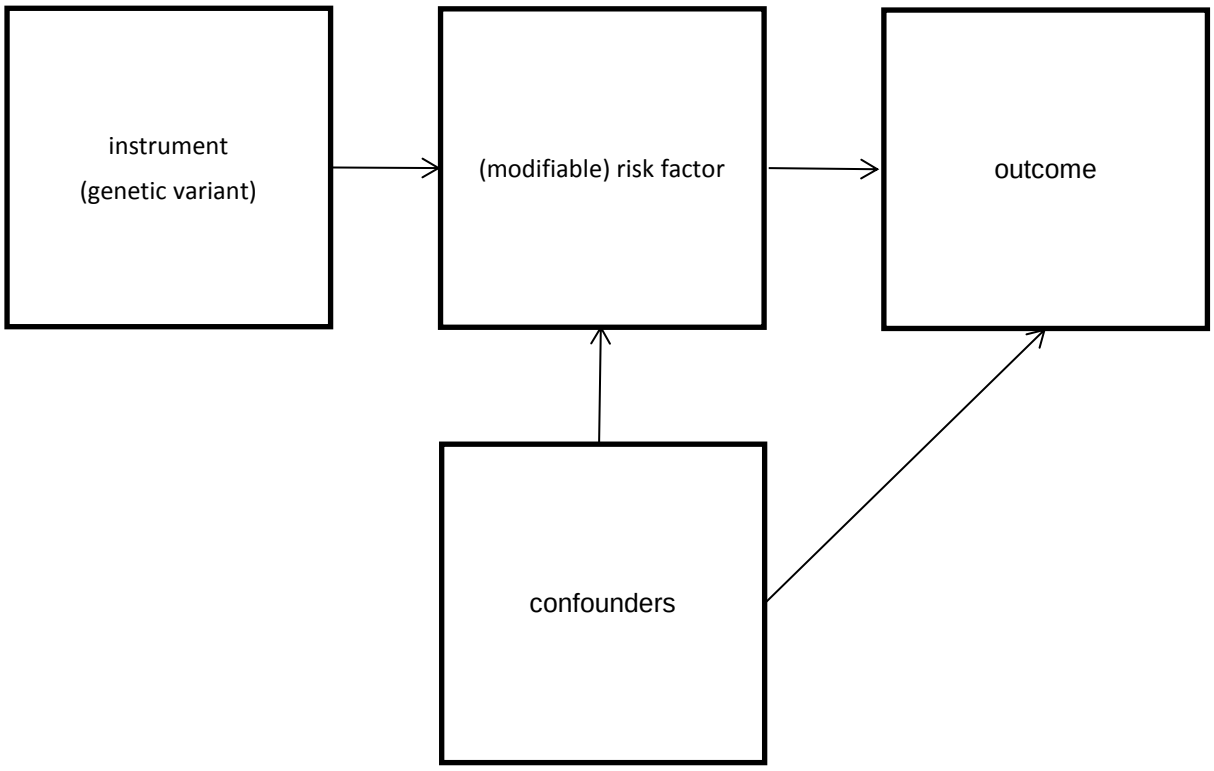


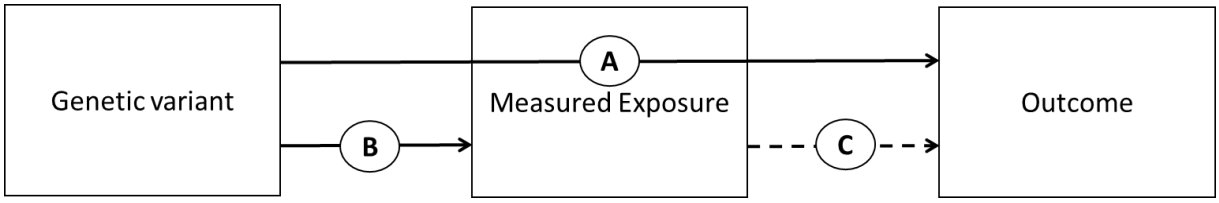
Figure from Davey Smith and Ebrahim, 2005 [10]. Reproduced with permission.

Figure 2. Principles of Mendelian Randomization.



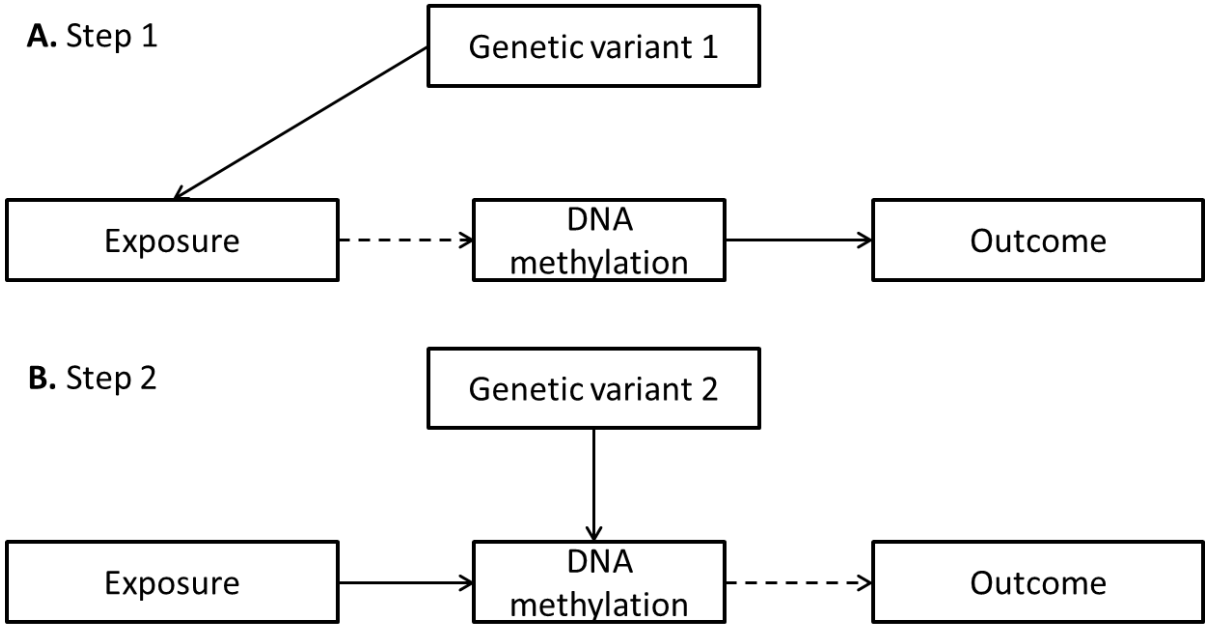
A directed acyclic graph illustrating the principles of Mendelian randomization is shown. Genotype is associated with the risk factor, but not the putative outcome or potential confounders.

Figure 3. Applications of Mendelian Randomization



The direct association between genetic variant and outcome (A) provides evidence for the existence of a causal relationship. The magnitude of the causal association between measured exposure and outcome (C) is calculated from the association between genetic variant and outcome (A) and the association between genetic variant and measured exposure (B).

Figure 4. Two-Step Epigenetic Mendelian Randomization



Dashed lines represent causal associations investigated at each step. In Step 1, a genetic variant associated with the exposure is used to investigate whether the exposure is causally associated with DNA methylation at a particular locus. In Step 2, a genetic variant associated with DNA methylation at the locus of interest is used to investigate whether DNA methylation is causally associated with the outcome. Diagram adapted from Relton and Davey Smith, 2012 [22]. Reproduced with permission.

Text Box 1. The CARTA Consortium

The consortium for Causal Analysis Research in Tobacco and Alcohol (CARTA; http://www.bris.ac.uk/expsych/research/brain/targ/research/collaborations/carta/) was established at the University of Bristol to investigate the causal effects of tobacco use, alcohol use and other lifestyle factors on health and sociodemographic outcomes using MR. CARTA includes over 30 studies, spanning nine countries, with a total sample size in excess of 150,000 – given the relatively small effects that individual genetic variants exert on exposures, MR generally requires very large sample sizes. CARTA has completed five initial analyses, investigating the impact of cigarette smoking on depression and anxiety, regional adiposity, blood pressure and heart rate, serum vitamin D levels and income. The genetic variant used as a proxy for this exposure was rs16969968, a genetic variant which is robustly associated with smoking heaviness in smokers [1-3,32,41,42]. Results of these initial analyses are currently in preparation.